Определение цены для нового товара

Senior Data Engineer
Селезнев Артем

 /NameArtem  /seleznev.artem.info  /seleznev-artem

Определение цены для нового товара

Senior Data Engineer
Селезнев Артем

БИЗНЕС ПОТРЕБНОСТЬ?

ПРОИЗВОДИТЕЛЬ	УСЛУГИ	ONLINE - X
?	?	?

БИЗНЕС ПОТРЕБНОСТЬ?

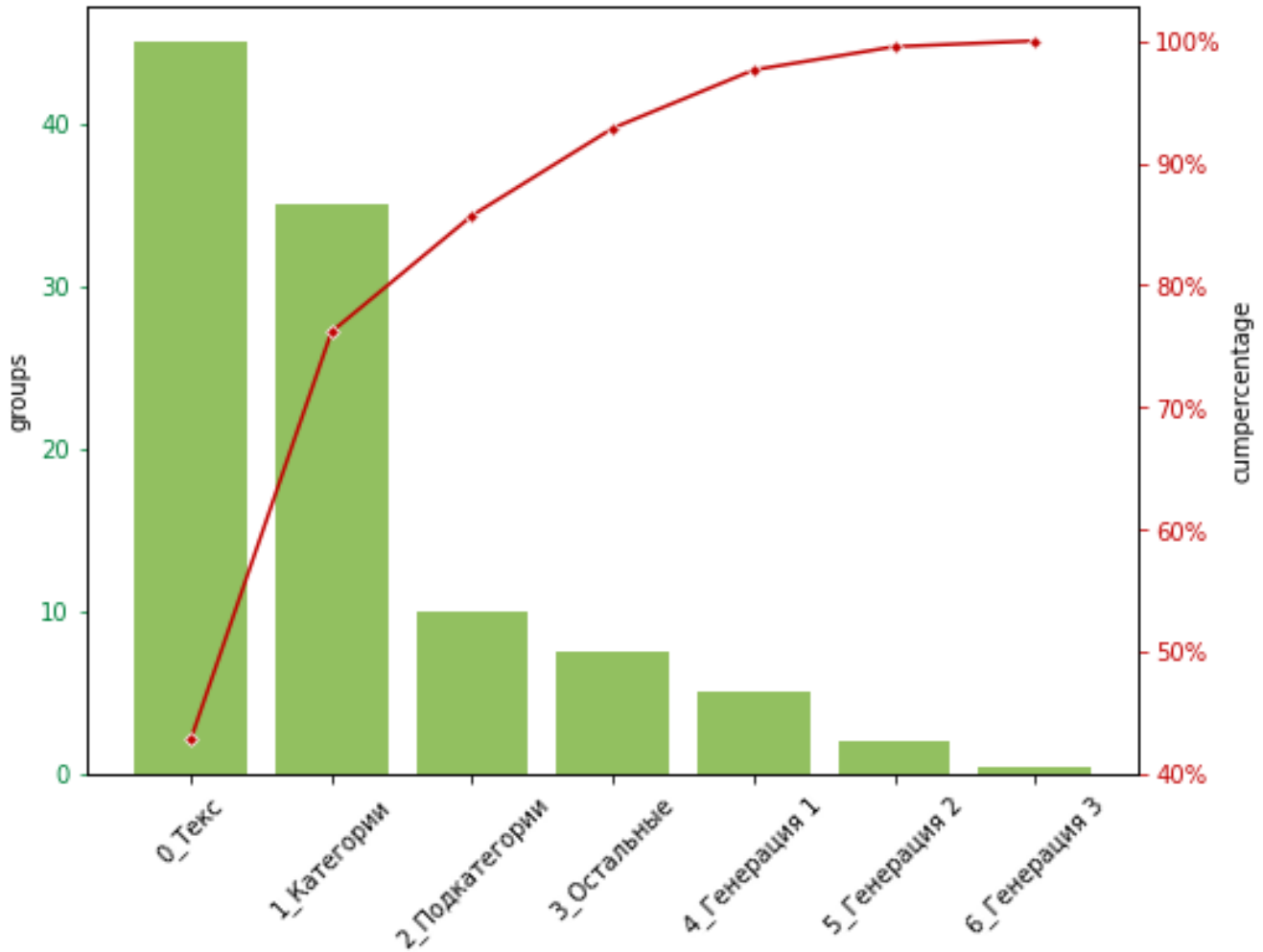
ПРОИЗВОДИТЕЛЬ	УСЛУГИ	ONLINE - X
X	+-	+

ДАННЫЕ ДЛЯ РАБОТЫ

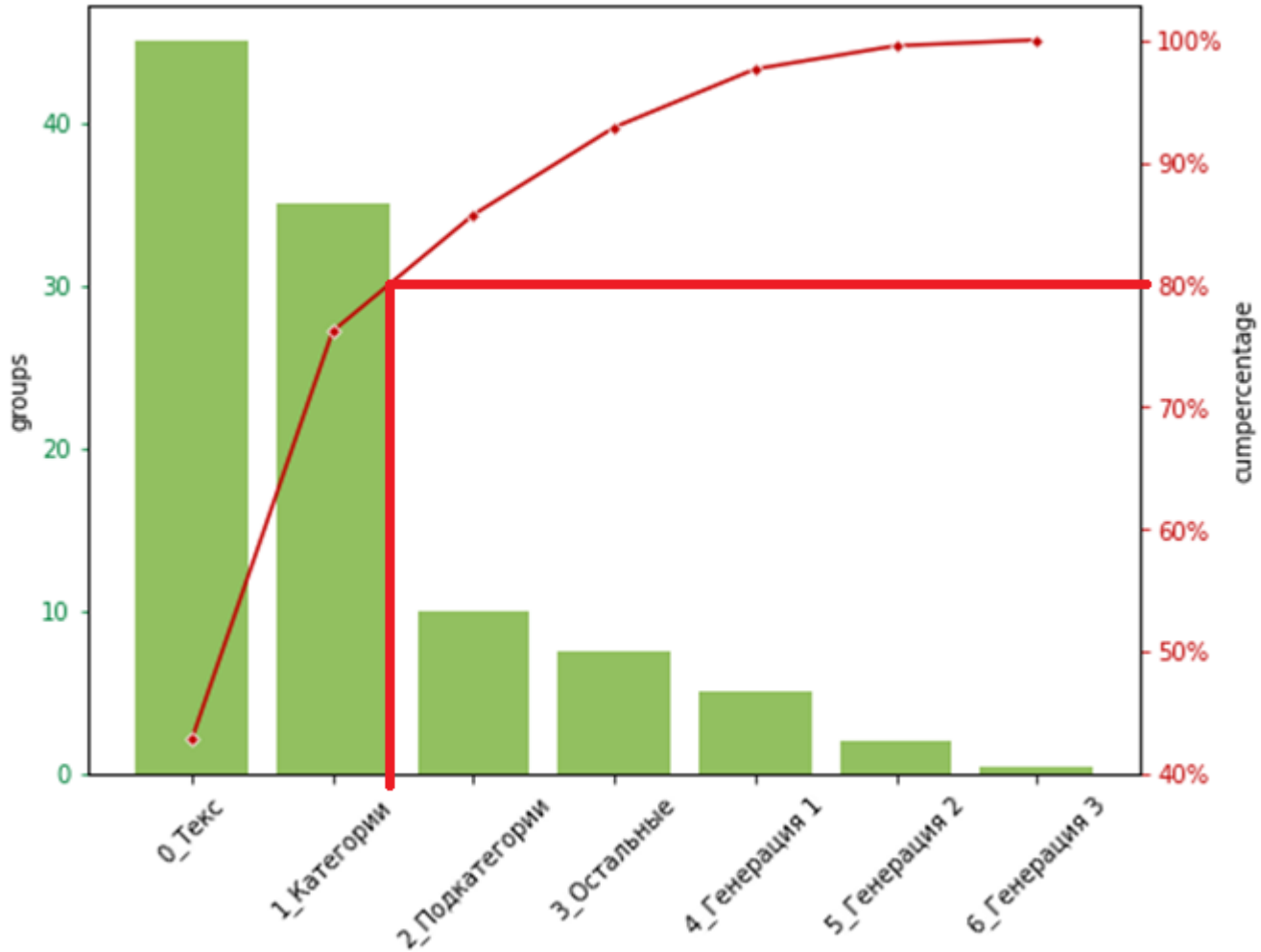
ПОЛУЧАЕМ ДАННЫЕ **И В БОЙ?**

«КАК СЪЕСТЬ ПИРОГ»

PARETO


groups
cumpercentage
0
10
20
30
40
40%
50%
60%
70%
80%
90%
100%
0_Текс
1_Категории
2_Подкатегории
3_Остальные
4_Генерация 1
5_Генерация 2
6_Генерация 3

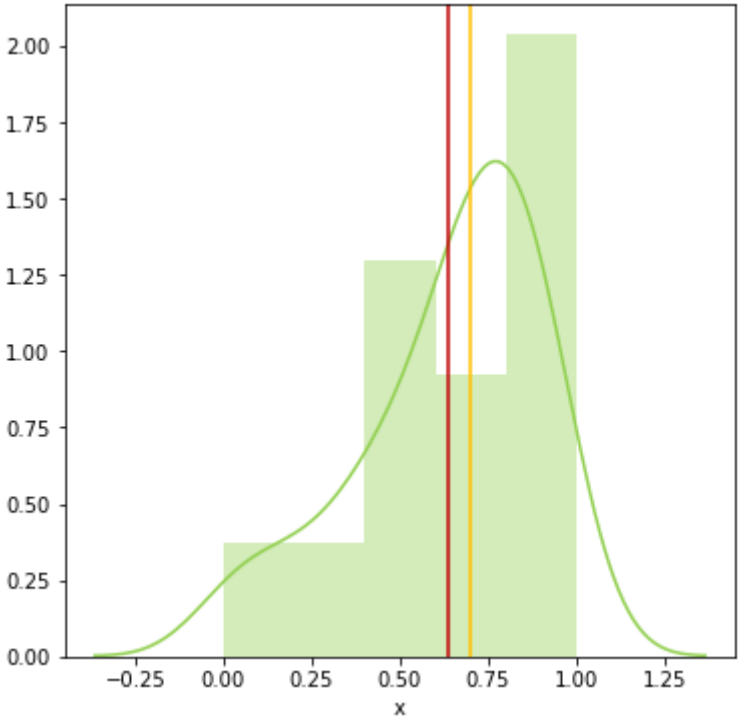
PARETO


groups
cumpercentage
0
10
20
30
40
40%
50%
60%
70%
80%
90%
100%
0_Текс
1_Категории
2_Подкатегории
3_Остальные
4_Генерация 1
5_Генерация 2
6_Генерация 3

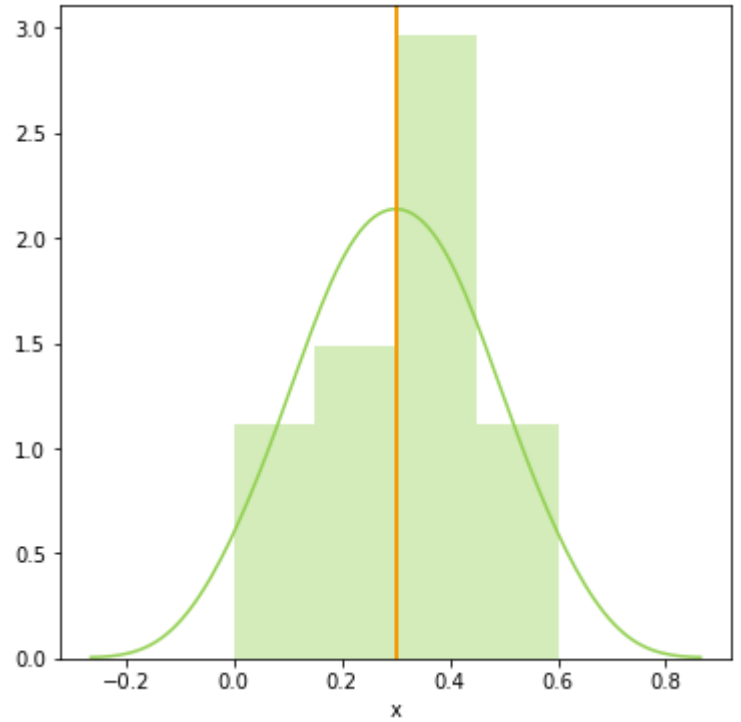
ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН (почему цена изменяется?)

- Количество поставки
- Качество товара
- Бренд
- Фабричное / ручное изделие
- Условия взаимодействия

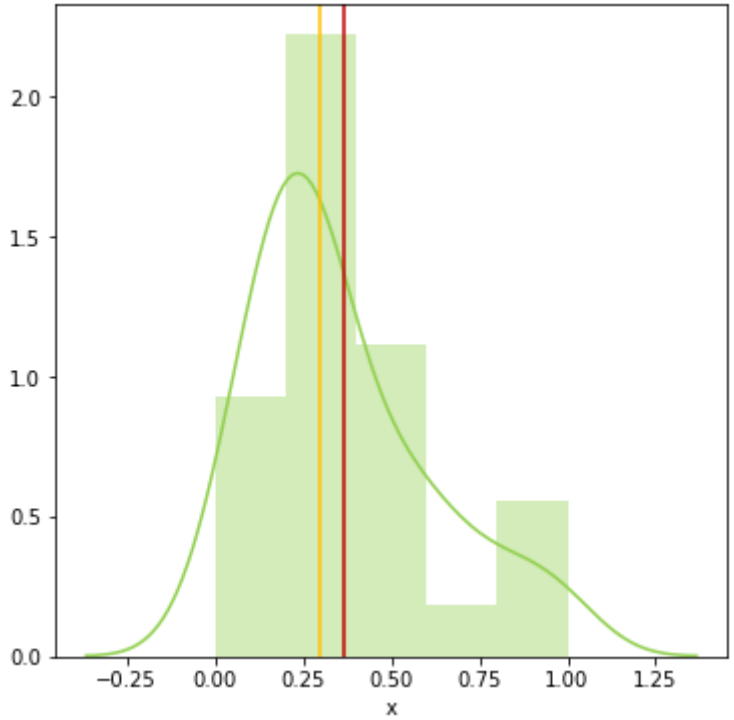
ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН
skewness / kurtosis



1

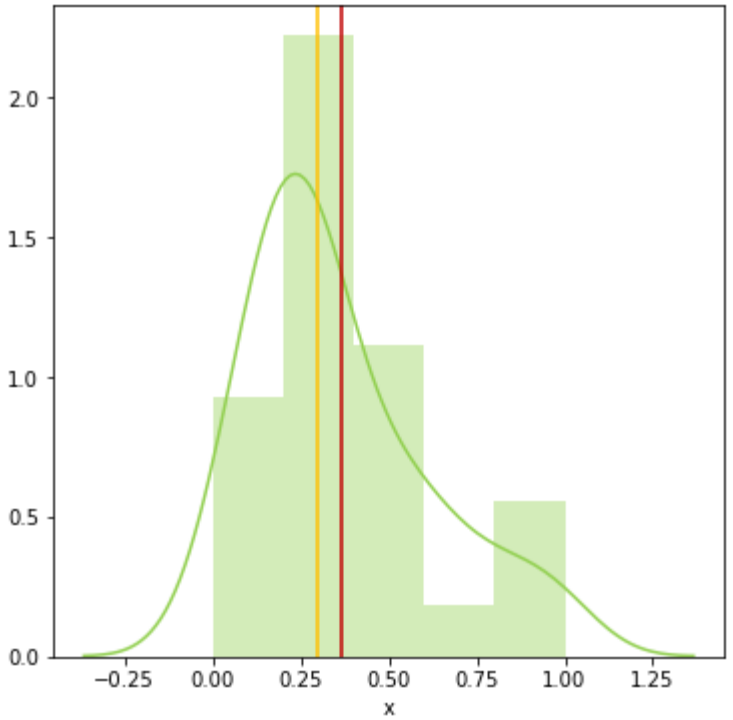


2



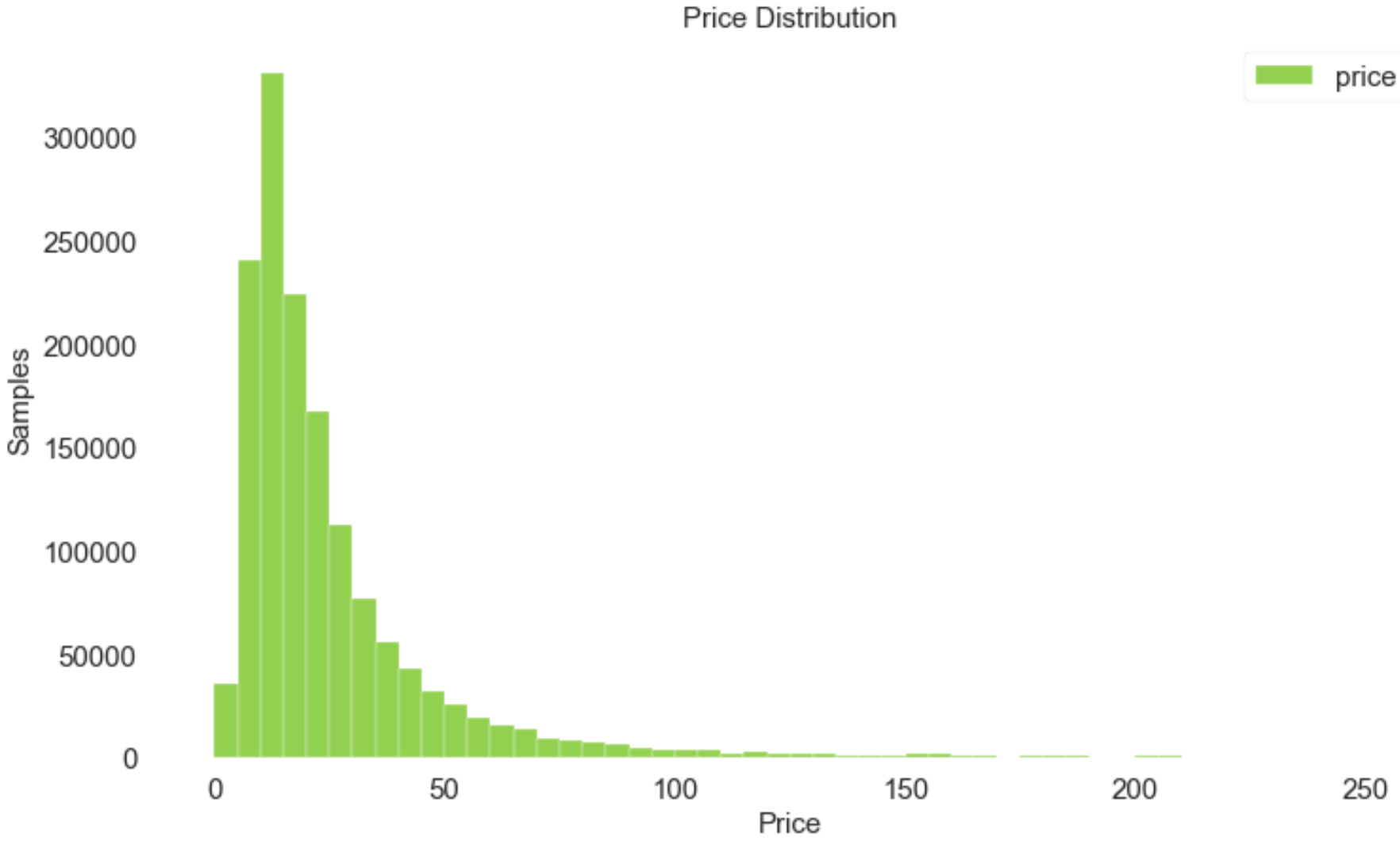
3

ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН

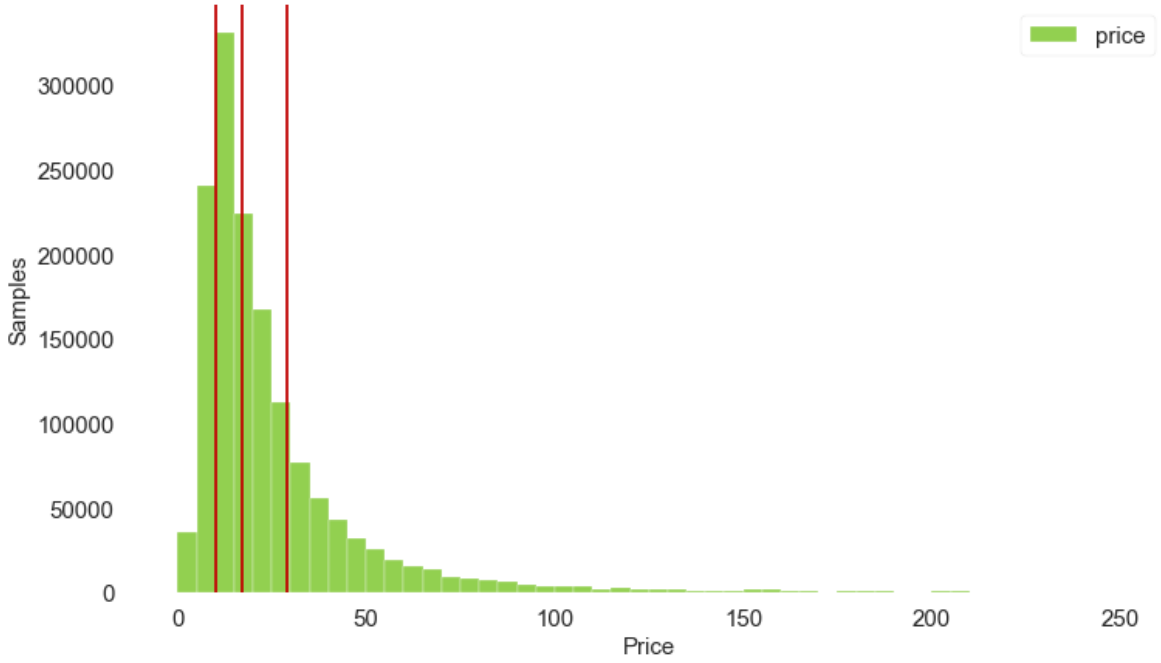


- Средняя стоимость товара **26 USD**
- Медиана **17 USD**
- Максимальная цена товара **2000 USD**

ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН

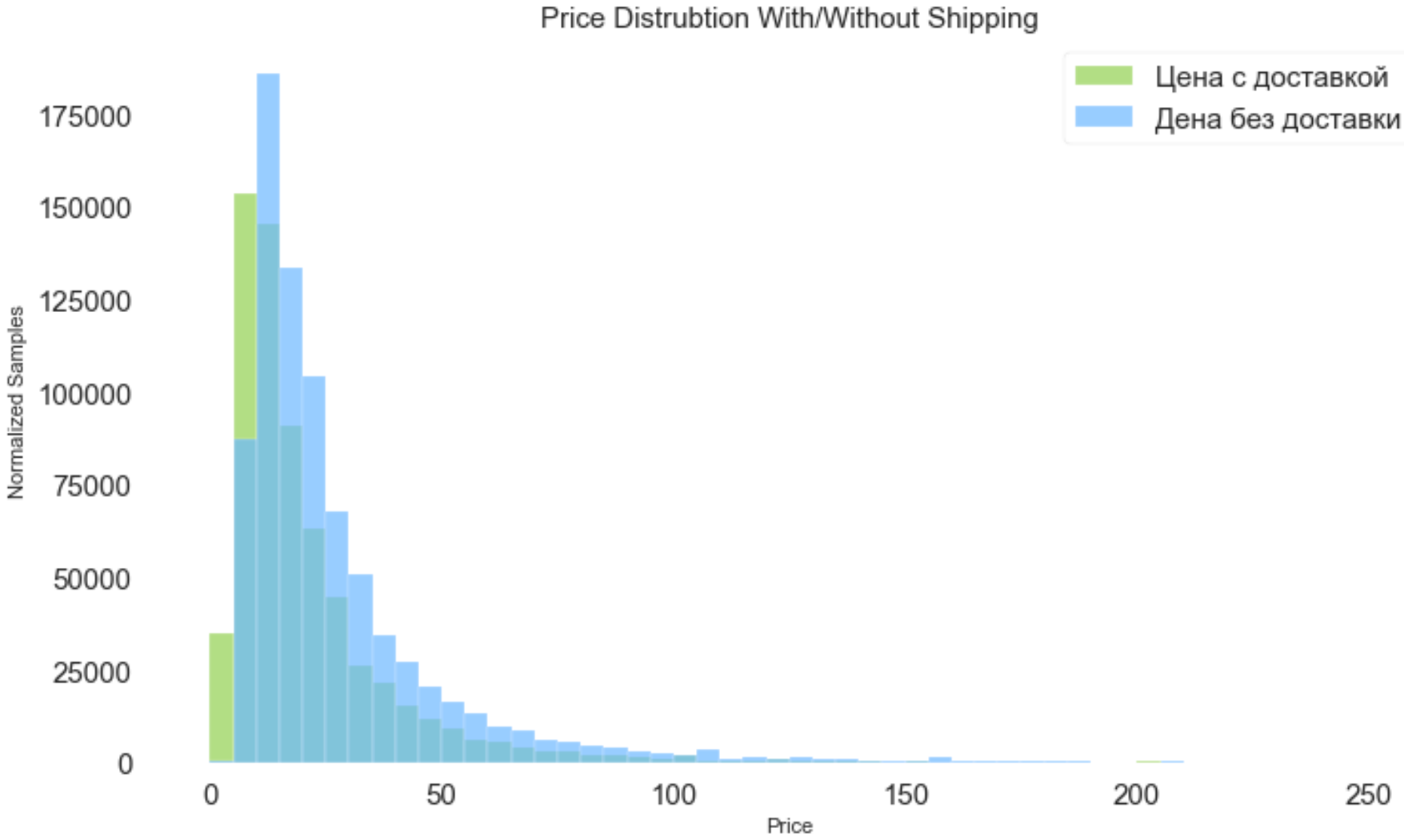


ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН



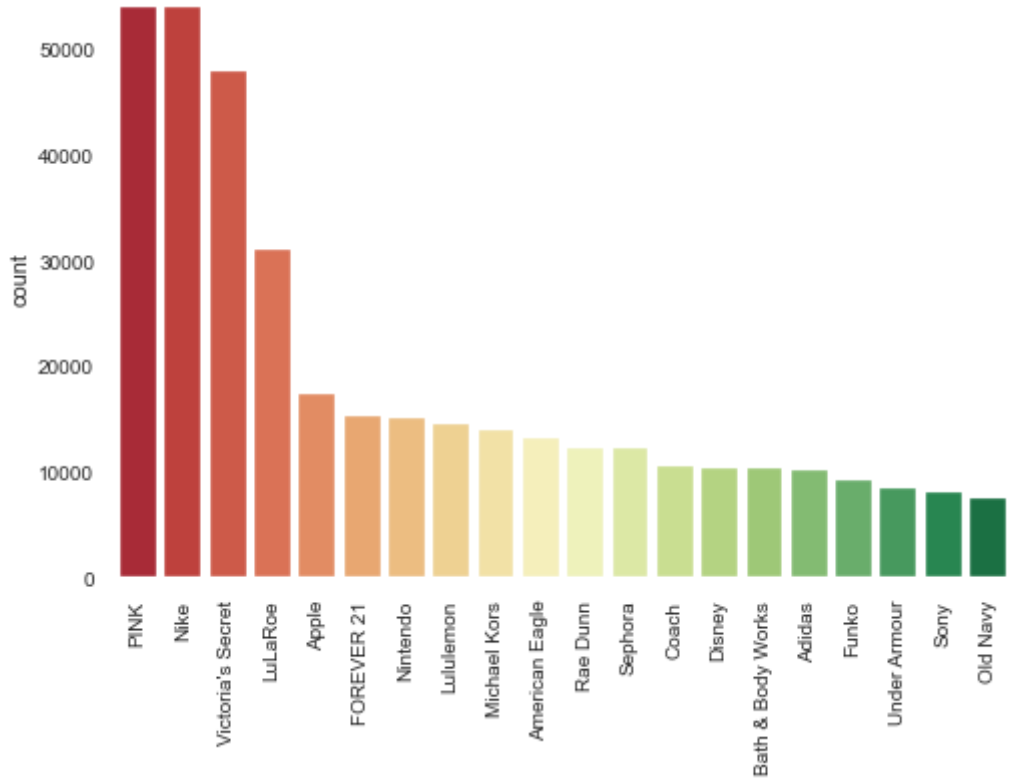
price_bin	count	mean	std	min	25%	50%	75%	max
q1	375615.0	7.715192	2.077888	3.0	6.0	8.0	10.0	10.0
q2	378177.0	13.842940	1.794584	10.5	12.0	14.0	15.0	17.0
q3	359743.0	22.555694	3.337832	17.5	20.0	22.0	25.0	29.0
q4	368123.0	63.527701	63.508250	29.5	35.0	45.0	66.0	2000.0

ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН



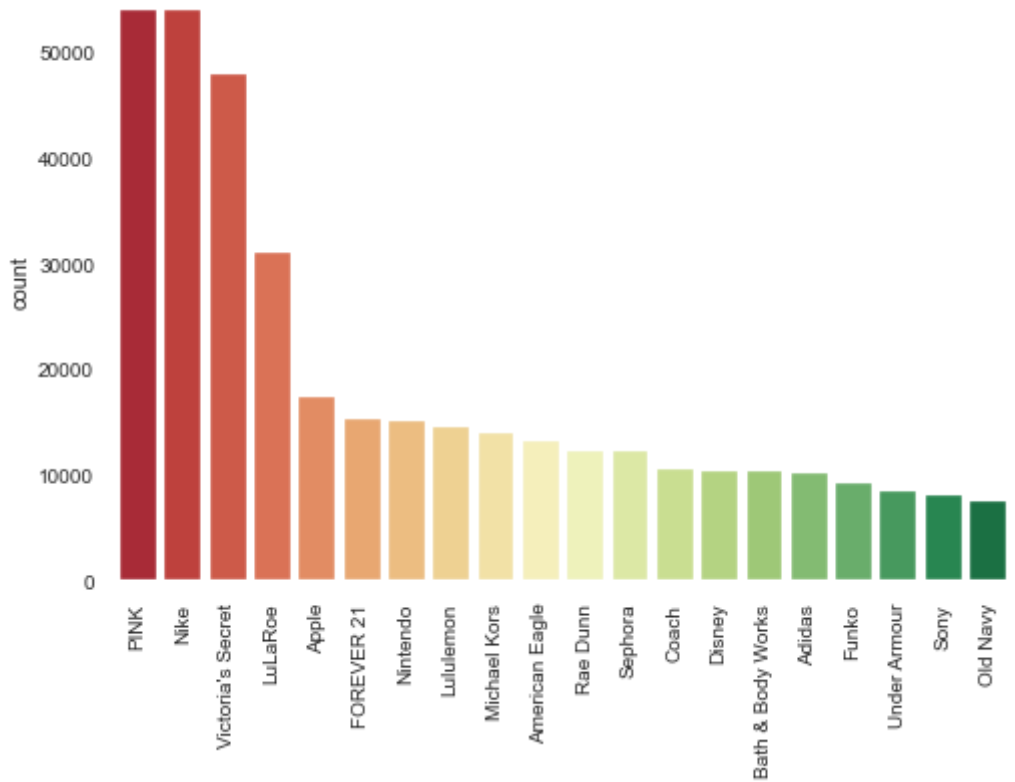
ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН

- Топ самых дорогих (max)

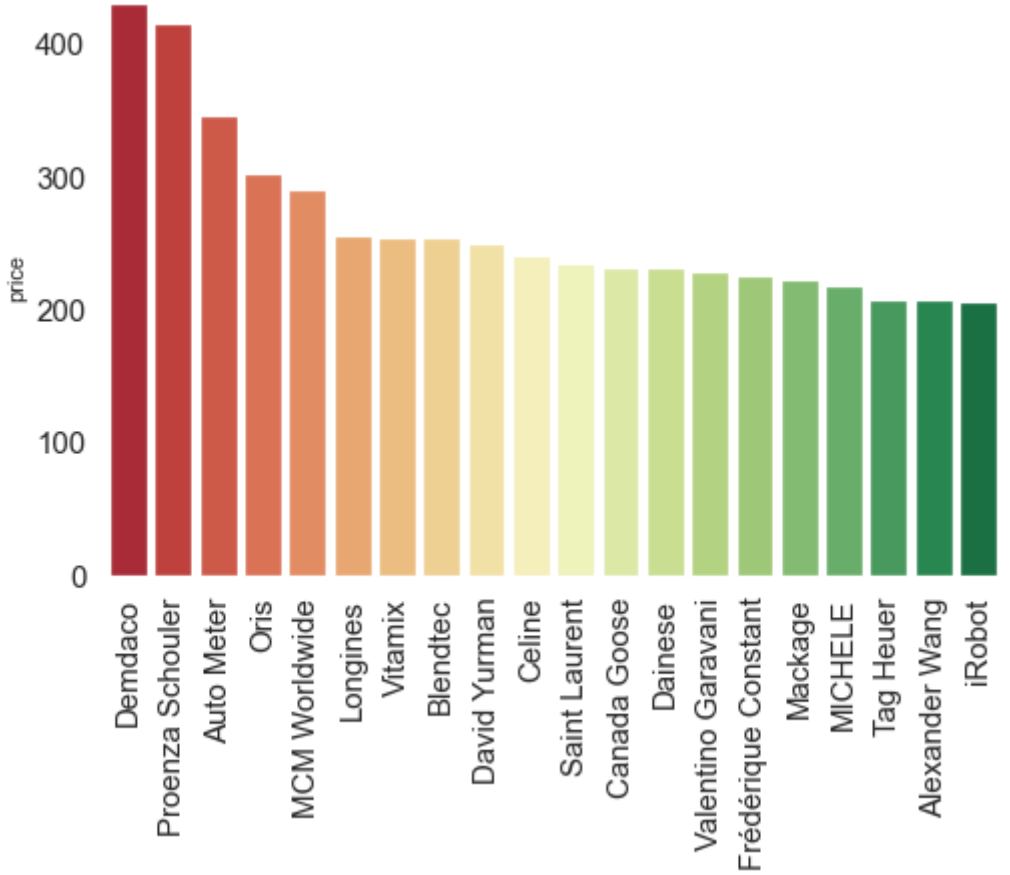


ИЗУЧАЕМ РАСПРЕДЕЛЕНИЕ ЦЕН

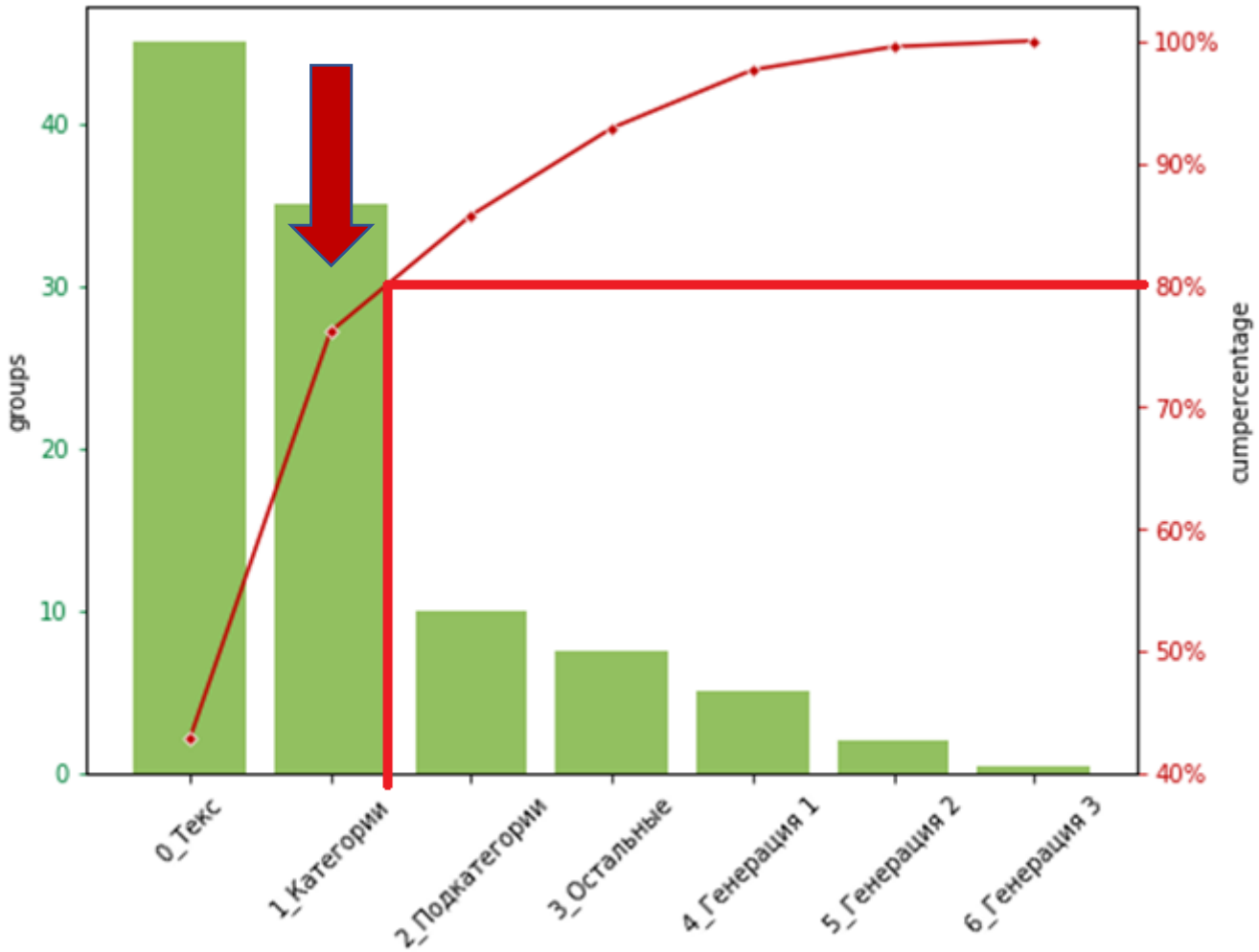
- Топ самых дорогих (max)



- Топ самых дорогих (avg)



КАТЕГОРИИ В ДАННЫЕ



	category_main	category_sub1	category_sub2	price
0	Men	Tops	T-shirts	10.0
1	Electronics	Computers & Tablets	Components & Parts	52.0
2	Women	Tops & Blouses	Blouse	10.0
3	Home	Home Décor	Home Décor Accents	35.0
4	Women	Jewelry	Necklaces	44.0

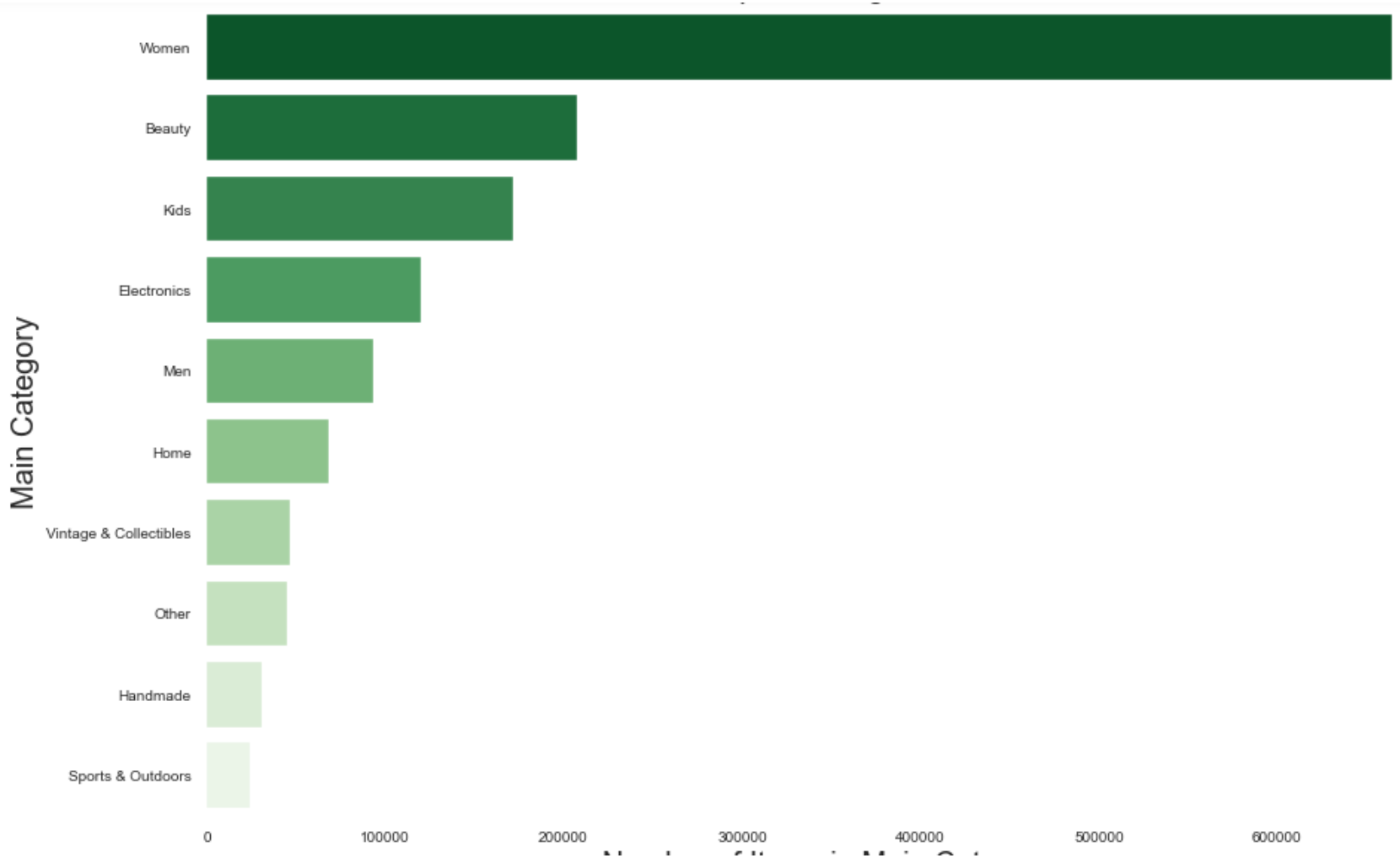
ГИПОТЕЗЫ

- Можем мы определить потребителя товара и разбить категорию на принадлежность к полу?

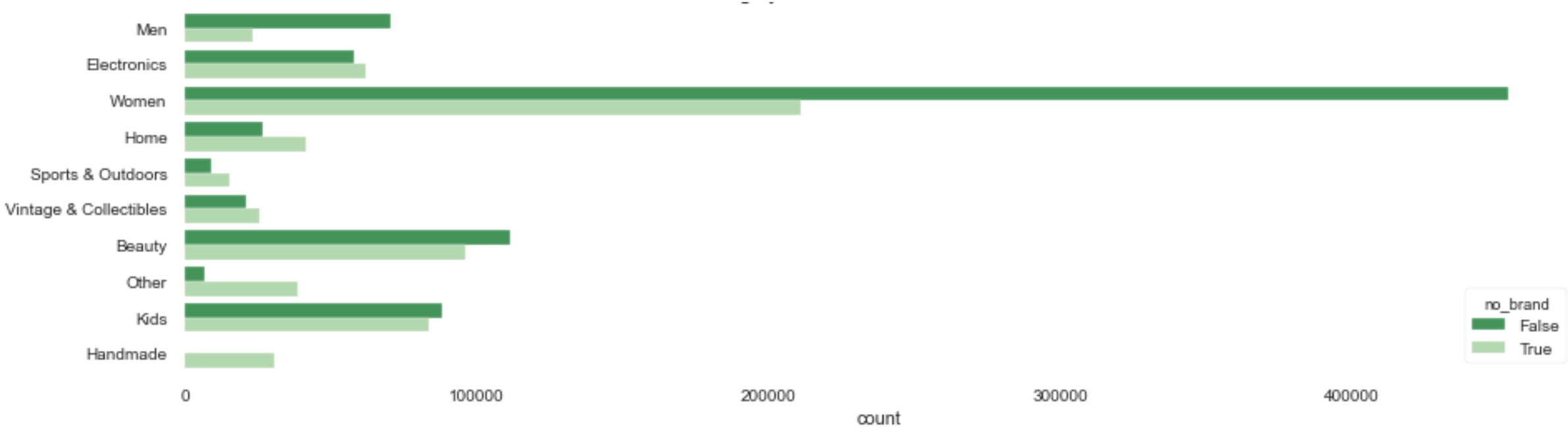
- Можем мы разбить категории с брендом / без бренда и спроецировать это по полу.
  Будет ли это значимо?

- Можем ли создать признак по возрасту и разбить его на категории?

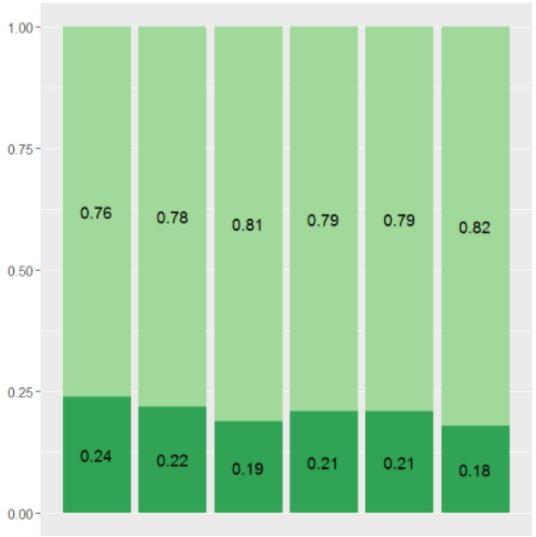
ГИПОТЕЗЫ


Main Category
Women
Beauty
Kids
Electronics
Men
Home
Vintage & Collectibles
Other
Handmade
Sports & Outdoors
0
100000
200000
300000
400000
500000
600000

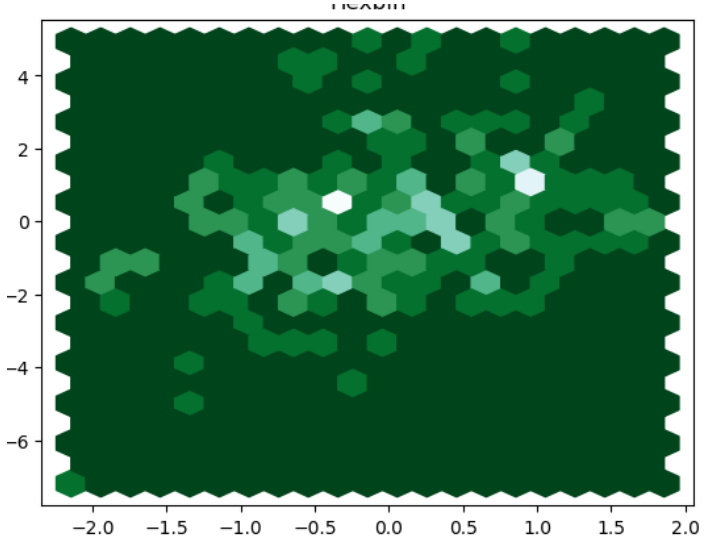
ГИПОТЕЗЫ


Men
Electronics
Women
Home
Sports & Outdoors
Vintage & Collectibles
Beauty
Other
Kids
Handmade
0
100000
200000
300000
400000
count
no_brand
False
True

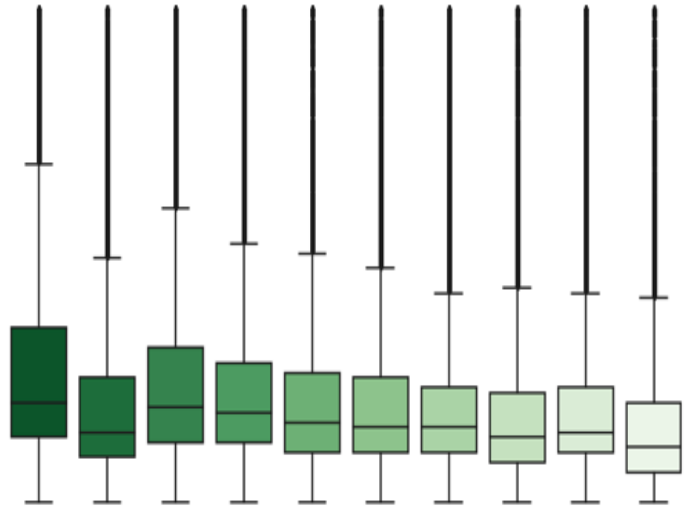
ВИЗУАЛИЗИРУЙ ПРАВИЛЬНО



1

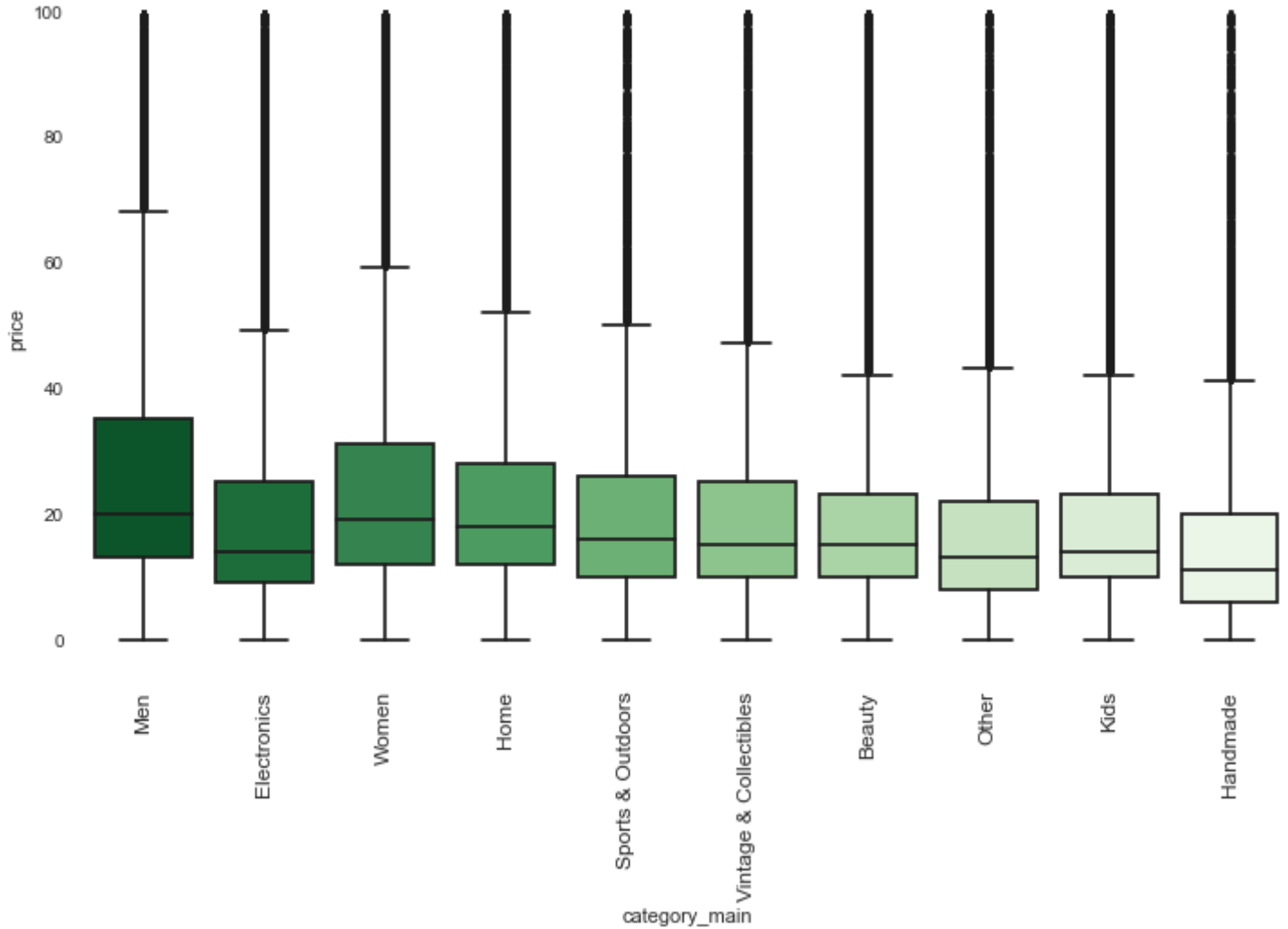


2



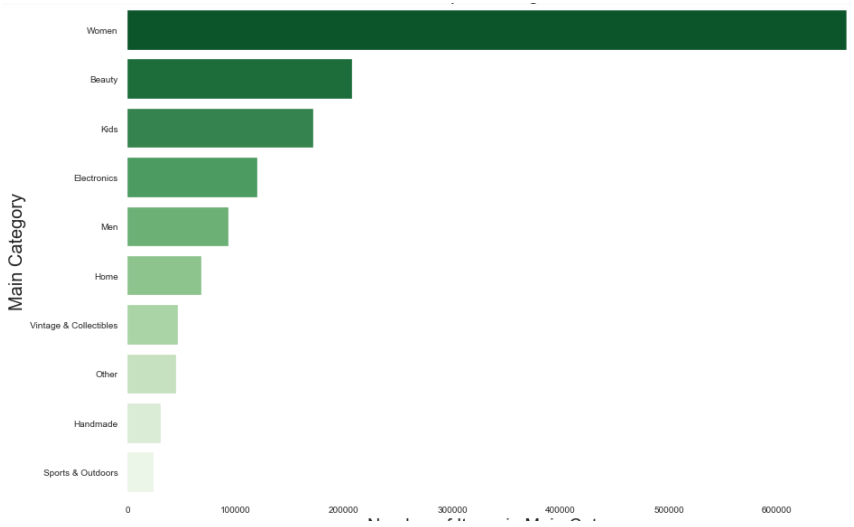
3

ВИЗУАЛИЗИРУЙ ПРАВИЛЬНО



20 -> 80 -> OHE

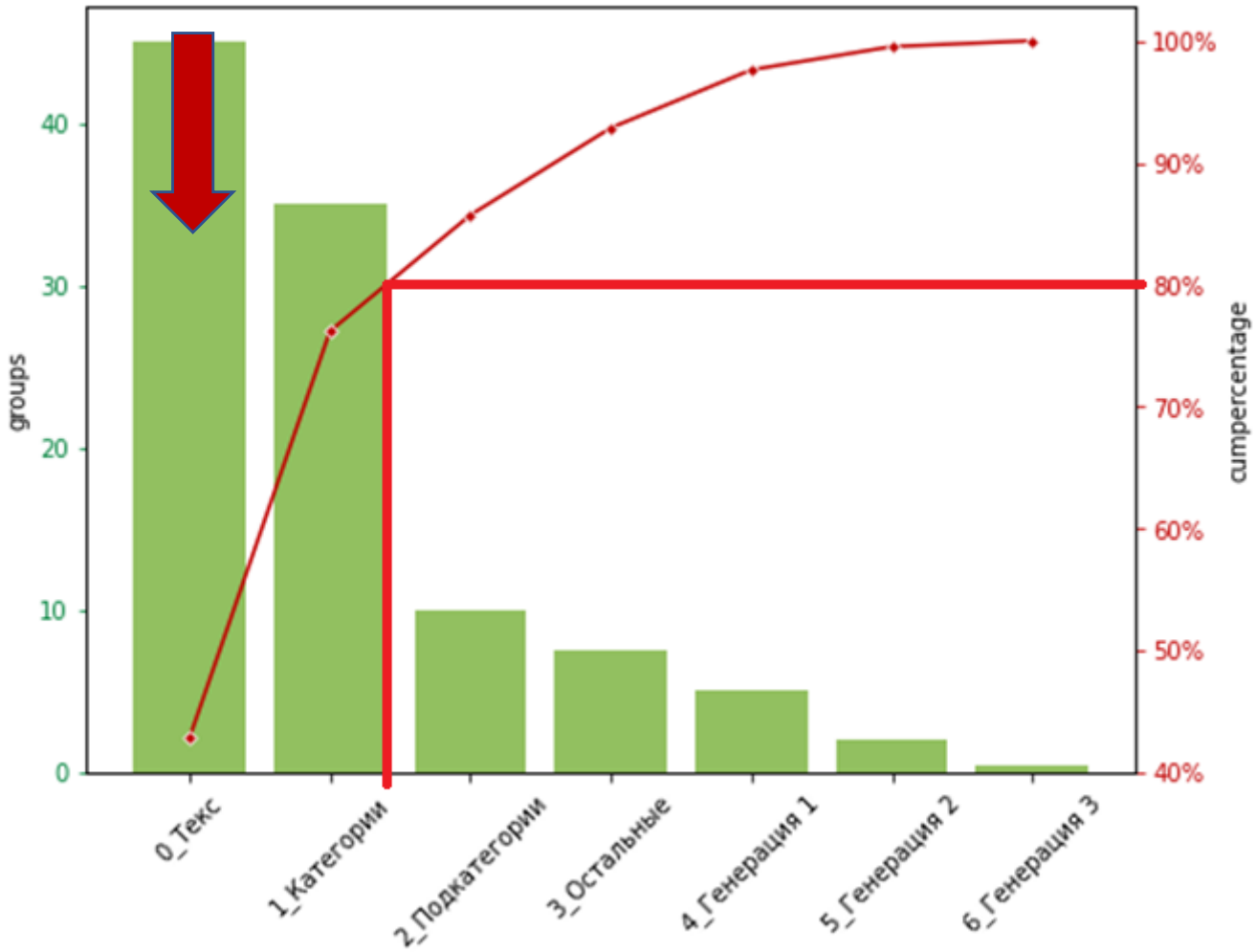
Top 20



One Hot Encoding

	Women	Beauty	Kids	Others
0	1	0	0	0
1	0	0	1	0
2	1	0	0	0
3	0	1	0	0
4	0	1	0	0
5	0	0	0	1

ТЕКСТ В ДАННЫЕ



item_description
Removable straps to make strapless Size: 34 B Perfect condition
Great Harry Potter Shirt! "Hogwarts, School of Witchcraft and Wizardry"! Women's
Brand new black and white ribbed mock neck bodysuit
Purple and Paisley Victoria's Secret Tankini Size Large. Worn a handful of times. E
[rm] for the set both in perfect condition no holes or stains like new medium in mer
Lace, says size small but fits medium perfectly too. Never used. Super cute for all
Little mermaid handmade dress never worn size 2t
Used once or twice, still in great shape.
There is 2 of each one that you see! So 2 red 2 orange and 2 of the big red and or
New with tag, red with sparkle. Firm price, no free shipping.

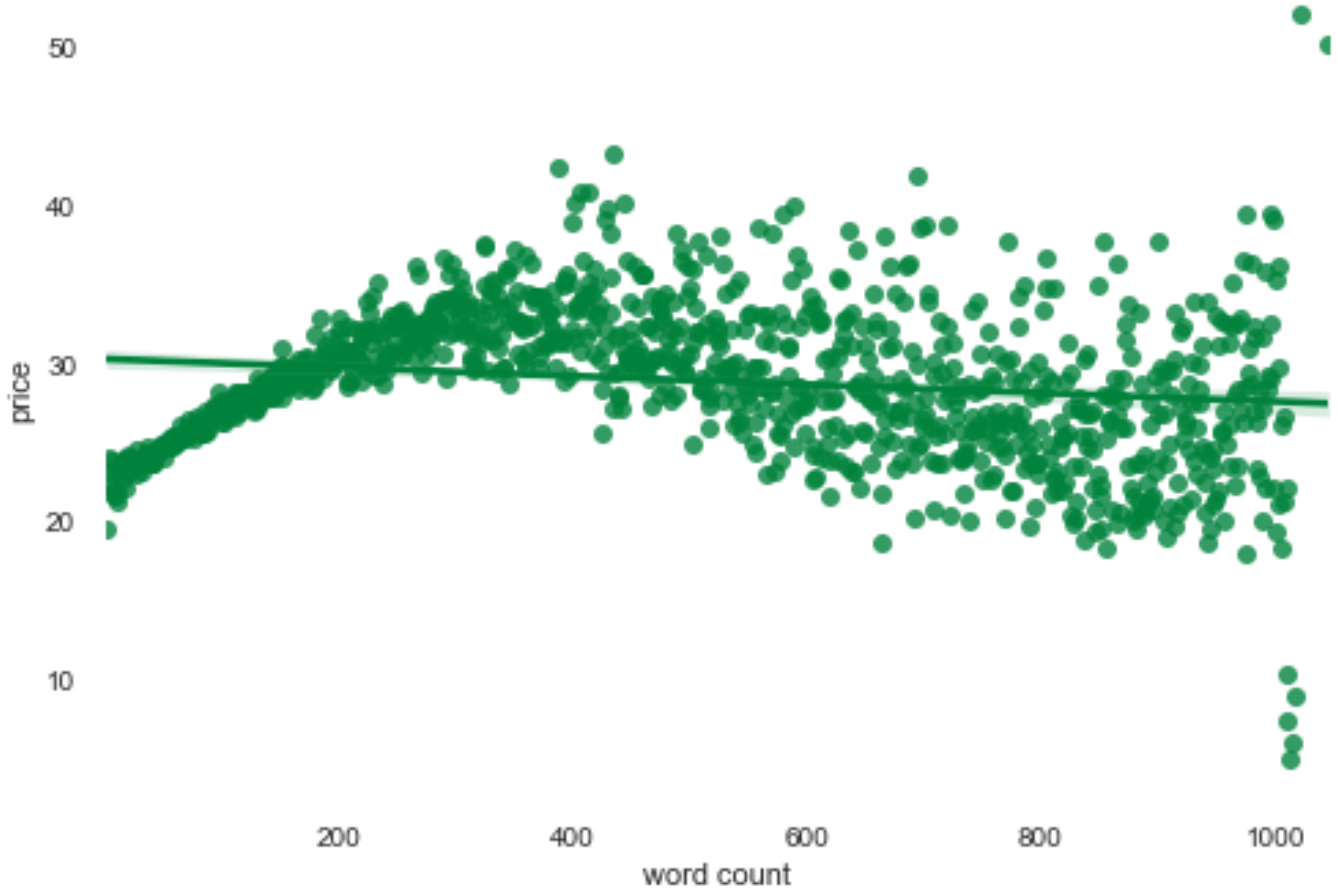
НОРМАЛИЗАЦИЯ ТЕКСТА

НОРМАЛИЗАЦИЯ ТЕКСТА

- Удаление стоп – слов
- Лемматизация
- Обработка от знаков
- Поиск N-грамм
- Ещё варианты?

item_description
description yet
keyboard great condition works like came box p...
adorable top hint lace key hole back pale pink...
new tags leather horses retail rm stand foot h...
complete certificate authenticity

WORD COUNT



TF / IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

TF / IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

- Частые и «не значимые» слова получают меньший «вес»
- Обнаружение важности документов за счет баланса «не значимых» и слов высокой важности

TF / IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

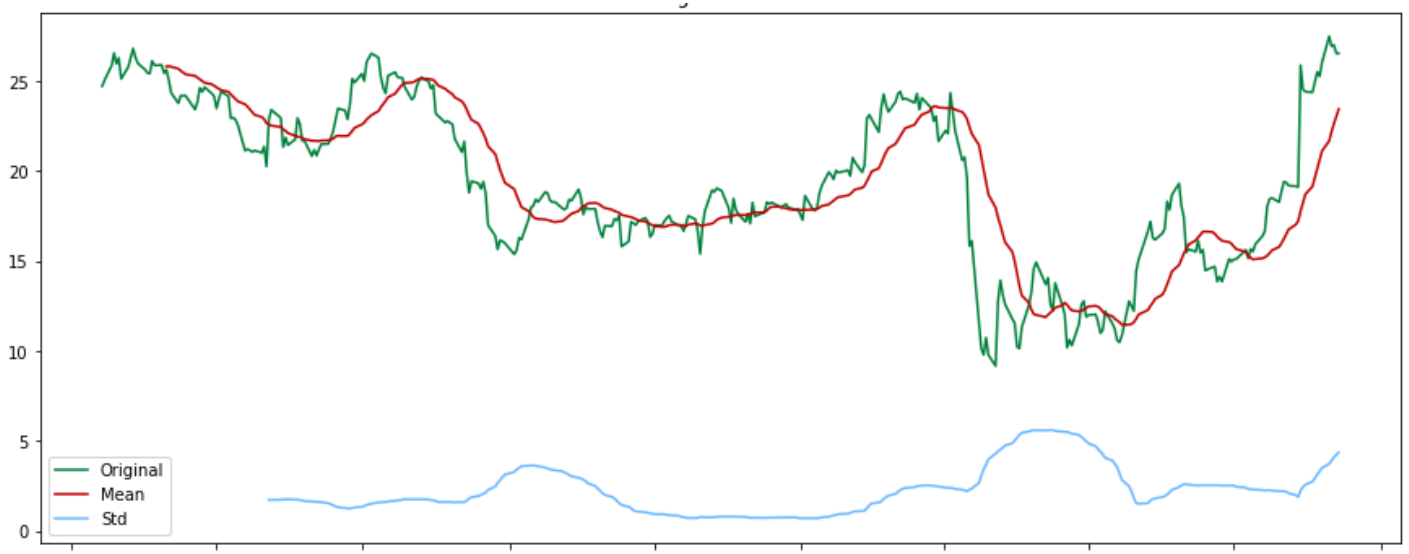
N = total number of documents

```
(98, 1306)    0.2365078875009367
(98, 602)     0.2279007964870353 7
(98, 1105)    0.1405902579454825 5
(98, 1303)    0.2073740859242166
(99, 845)     0.2361583444950507 4
```

ЧТО ЕЩЁ МОЖНО ДОБАВИТЬ?

ЧТО ЕЩЁ МОЖНО ДОБАВИТЬ?

- Учитывать изменение цен



ЧТО ЕЩЁ МОЖНО ДОБАВИТЬ?

- Учитывать изменение цен

Изменяемость на 10 дней	0.000000
Плавающее среднее на 50 дней	359.000000
Критическое значение (1%)	-3.448697
Критическое значение (5%)	-2.869625
Критическое значение (10%)	-2.571077

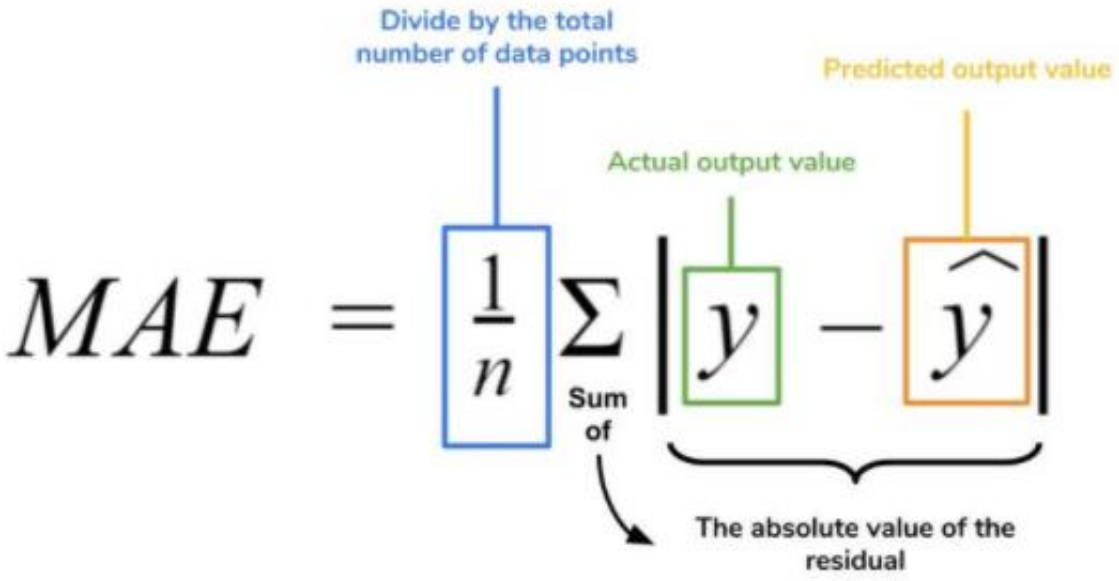
ВСЕ ГОТОВО К ОБУЧЕНИЮ?

МЕТРИКА + ЦЕЛЕВОЕ ЗНАЧЕНИЕ

МЕТРИКА

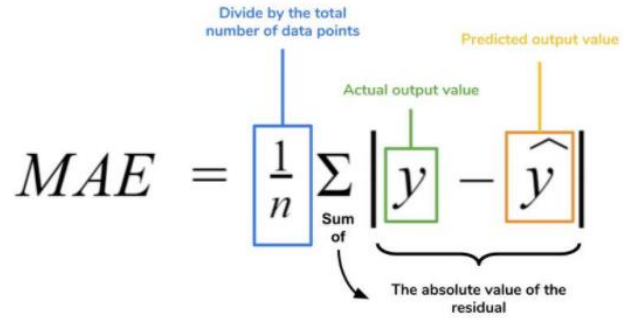
$$\text{Residual Error} = y - \hat{y}$$

МЕТРИКИ РЕГРЕССИИ



МЕТРИКИ РЕГРЕССИИ

Факт	Предсказание	Ошибка	Абсолютная ошибка
150	130	20	20
100	120	-20	20
60	70	-10	10
205	200	5	5
		Среднее значение	**13,75**



МЕТРИКИ РЕГРЕССИИ

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

МЕТРИКИ РЕГРЕССИИ

Факт	Предсказание	Ошибка	Квадрат ошибки
150	130	20	400
100	120	-20	400
60	70	-10	100
205	200	5	25
		Среднее значение	**231,25**

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

МЕТРИКИ РЕГРЕССИИ

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

МЕТРИКИ РЕГРЕССИИ

Факт	Предсказание	Ошибка	Квадрат ошибки
150	130	20	400
100	120	-20	400
60	70	-10	100
205	200	5	25
		Корень от среднего	**15,20690633**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

МЕТРИКИ РЕГРЕССИИ

$$\text{RMSLE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(y_i+1\right)-\left(\widehat{y+1}\right)\right)^2}$$

МЕТРИКИ РЕГРЕССИИ

Факт	Предсказание	Факт + 1	Предсказание + 1	Log Факта	Log Предсказания	Ошибка	Квадрат ошибки
150	130	151	131	2,176	2,113	0,062	0,003
100	120	101	121	2	2,079	-0,079	0,006
60	70	61	71	1,778	1,845	-0,066	0,004
205	200	206	201	2,311	2,301	0,0107	0,0005
						Корень от среднего	0,06068135

$$\text{RMSLE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(y_i+1\right)-\widehat{(y+1)}\right)^2}$$

ЭВОЛЮЦИЯ Y

0	10.0
1	52.0
2	10.0
3	35.0
4	44.0
	...
995	13.0
996	17.0
997	7.0
998	3.0
999	12.0

?

ЭВОЛЮЦИЯ Y

0	10.0
1	52.0
2	10.0
3	35.0
4	44.0
	...
995	13.0
996	17.0
997	7.0
998	3.0
999	12.0

?

0	2.397895
1	3.970292
2	2.397895
3	3.583519
4	3.806662
	...
995	2.639057
996	2.890372
997	2.079442
998	1.386294
999	2.564949

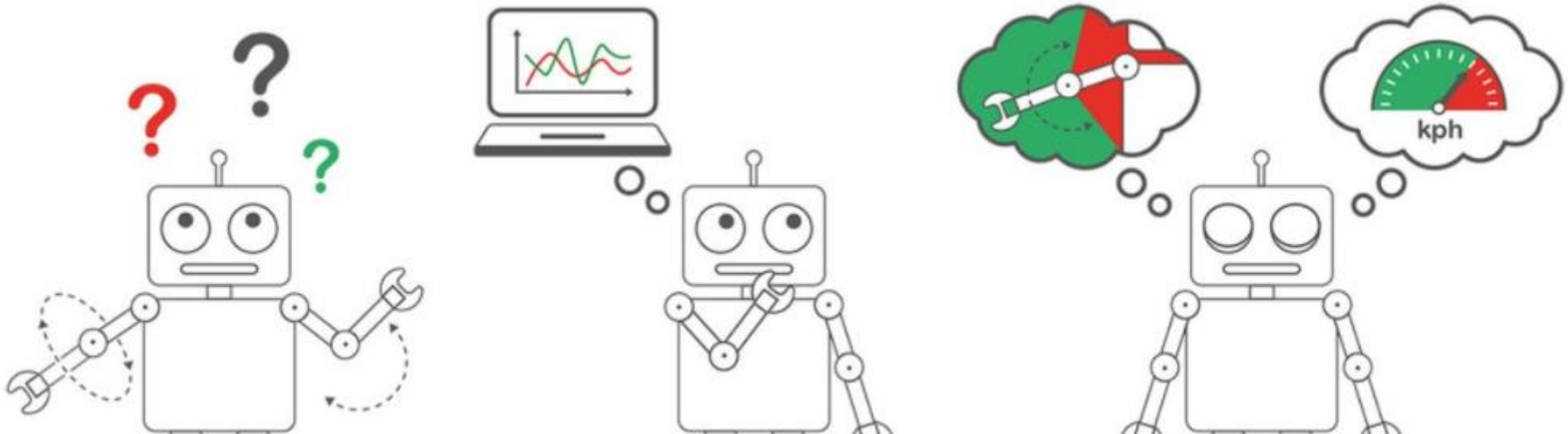
ЭВОЛЮЦИЯ Y

0	10.0
1	52.0
2	10.0
3	35.0
4	44.0
	...
995	13.0
996	17.0
997	7.0
998	3.0
999	12.0

Log(Y)

0	2.397895
1	3.970292
2	2.397895
3	3.583519
4	3.806662
	...
995	2.639057
996	2.890372
997	2.079442
998	1.386294
999	2.564949

УЧИМ АЛГОРИТМ



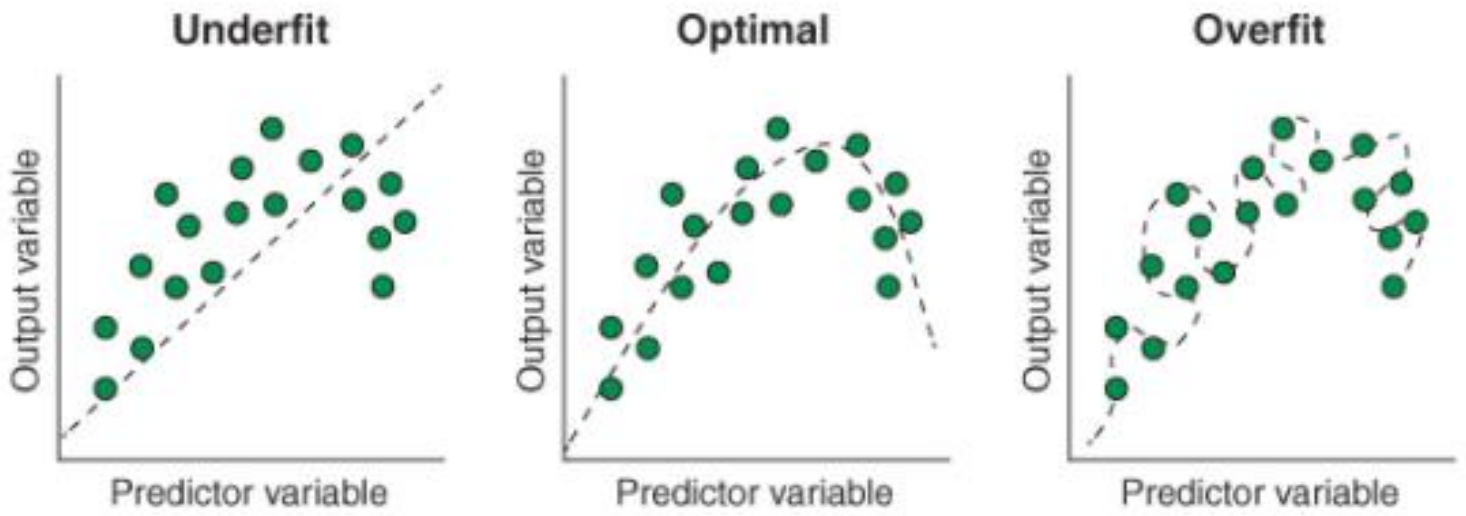
BASELINE

RIDGE REGRESSION	BOOSTING
?	?

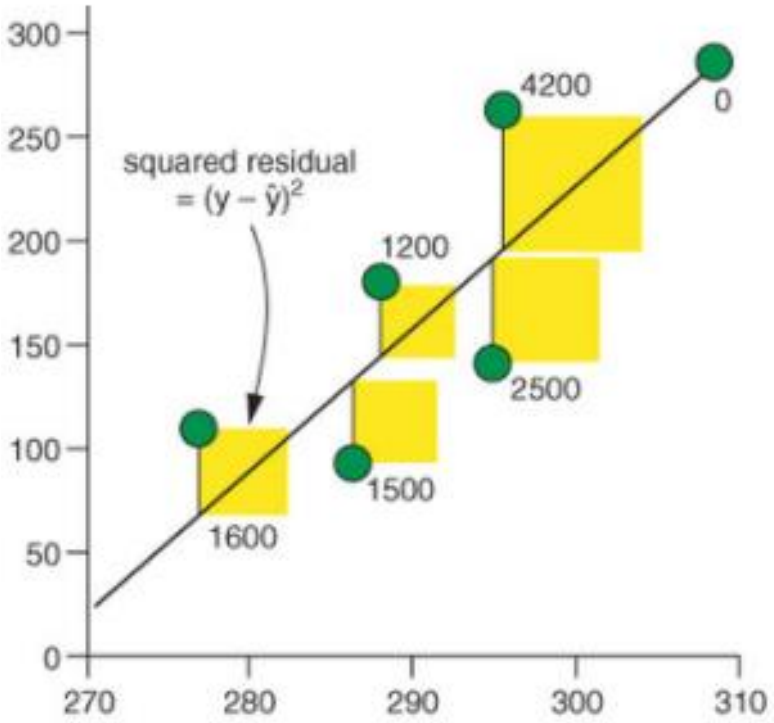
BASELINE

RIDGE REGRESSION	BOOSTING
	

REGRESSION


Underfit
Output variable
Predictor variable
Optimal
Output variable
Predictor variable
Overfit
Output variable
Predictor variable

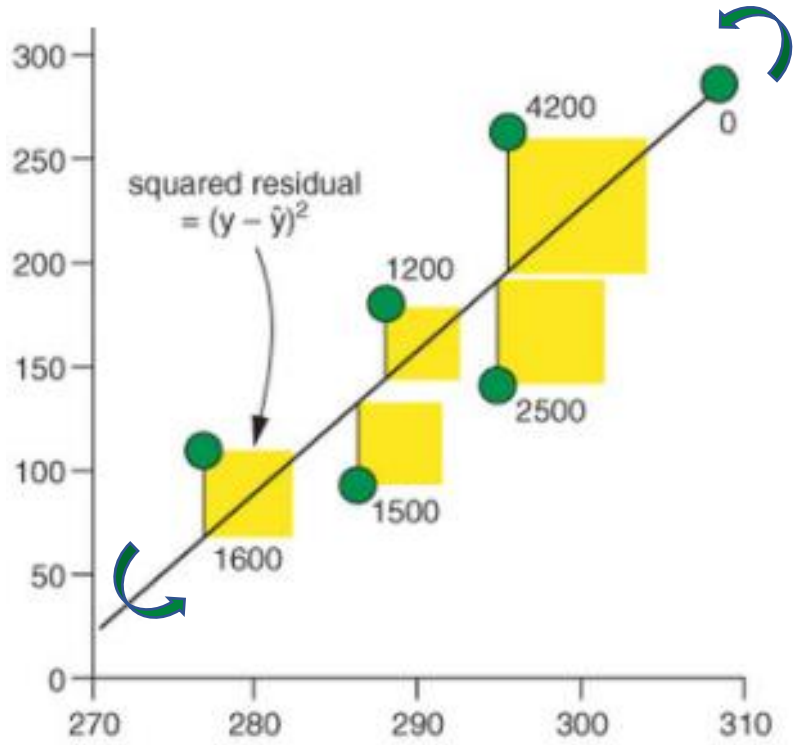
RIDGE REGRESSION



- Посчитать все потери сумме квадратов потерь

  sum = (fact – pred)^2

RIDGE REGRESSION



- Посчитать все потери сумме квадратов потерь

  sum = (fact – pred)^2

- Посчитать параметр (L2)

BOOSTING

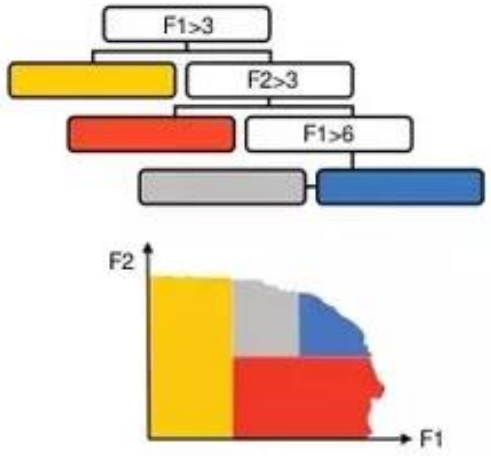
XGBoost	LGB	CatBoost
+ DropOut + Хороший из «коробки» - Нет работы с категориями	+ Работает с категориальными данными + Скорость работы + Мало ОЗУ	+ Работает с категориальными данными + Много настроек + Много «полезностей»

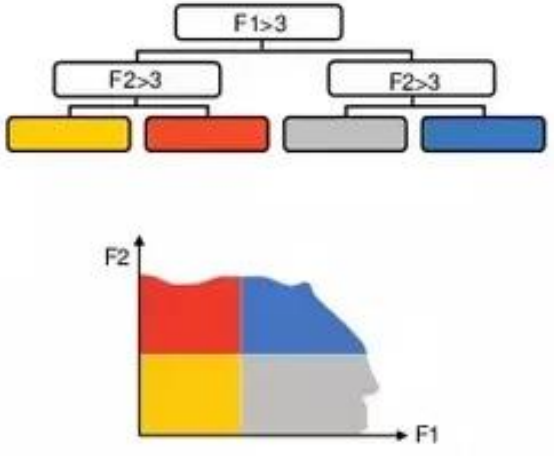
BOOSTING

CatBoost

LGB

XGBoost



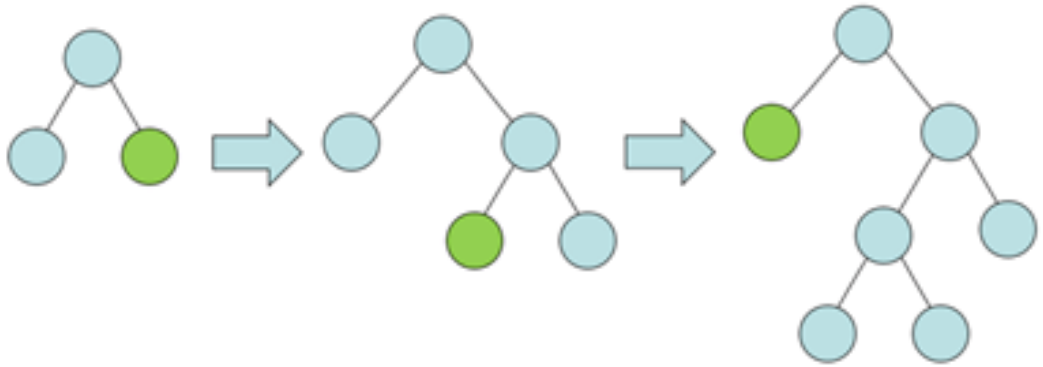


BOOSTING

CatBoost

LGB

XGBoost

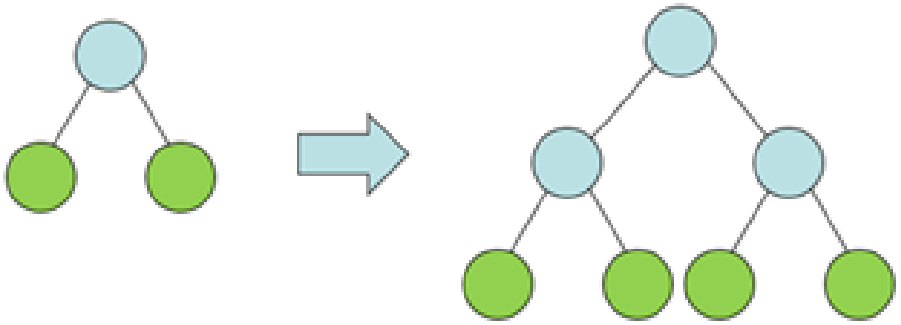


BOOSTING

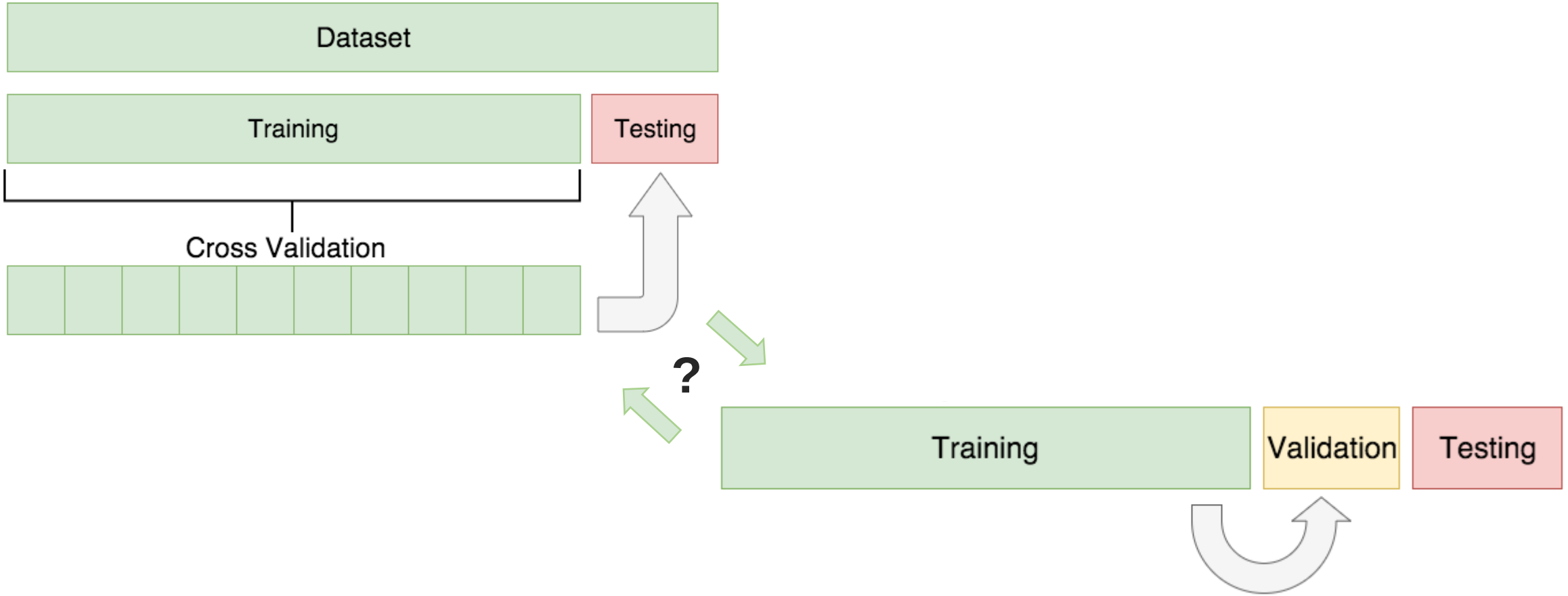
CatBoost

LGB

XGBoost



TRAIN | TEST | VALIDATE



TEST

RIDGE REGRESSION

BOOSTING

TEST

RIDGE REGRESSION

BOOSTING

The RMSLE of Ridge Regression is: **0.4829**

TEST

RIDGE REGRESSION

The RMSLE of Ridge Regression is: **0.4829**

BOOSTING

The RMSLE of LGBM is: **0.5406**

Полезные ссылки

Центр непрерывного образования ФКН ВШЭ

21 августа 19:00 вебинар «Зачем нужны алгоритмы и структуры данных?»
26 августа 19:00 вебинар «RFM-анализ: как выявить и удержать ключевых клиентов»

Программы профессиональной переподготовки:

- «Аналитик данных» Старт 8 сентября
- «Специалист по Data Science» Старт 23 сентября

Соцсети:

https://www.facebook.com/hsecs/

https://vk.com/cshse

https://t.me/fcs_hse